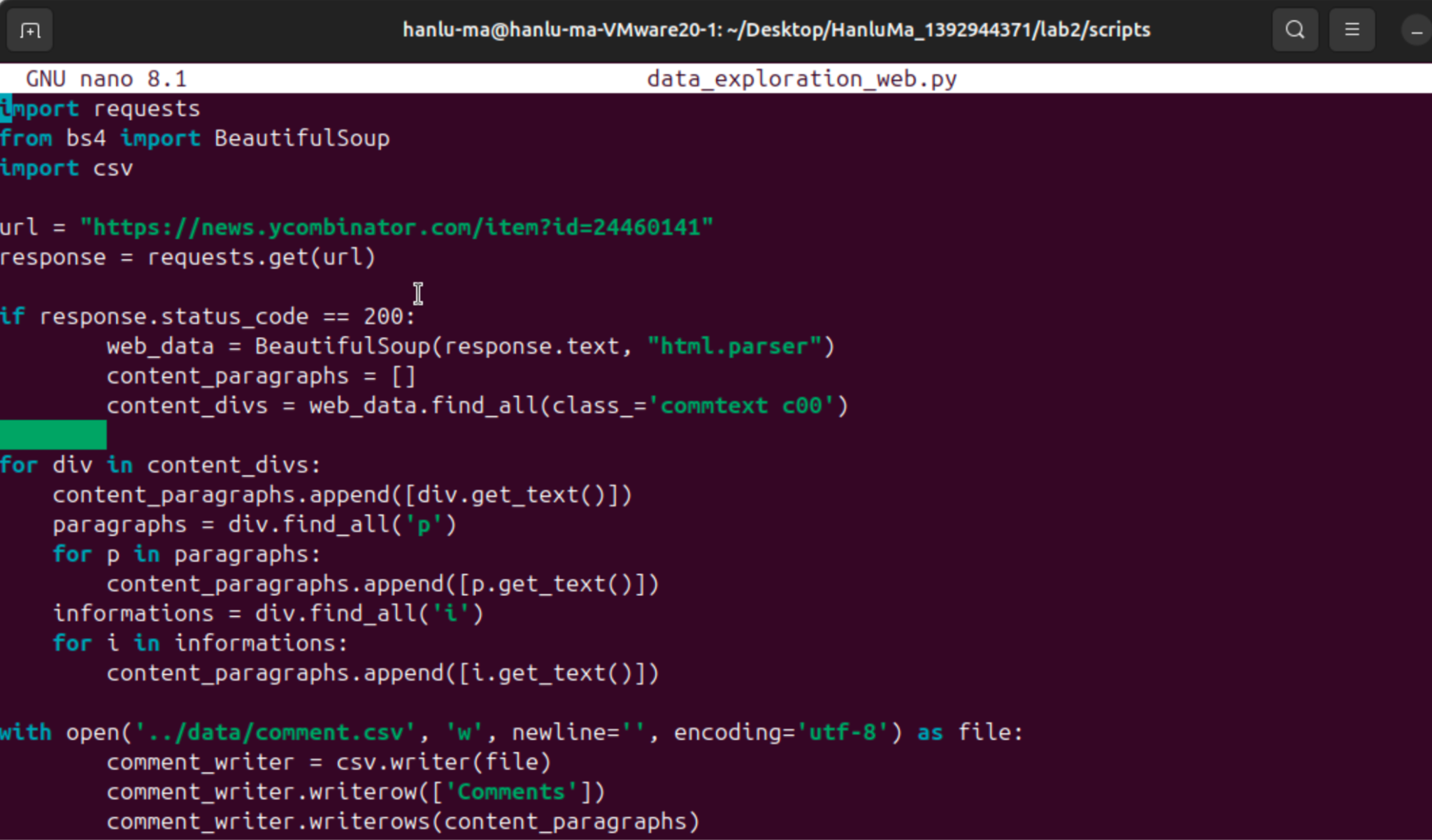
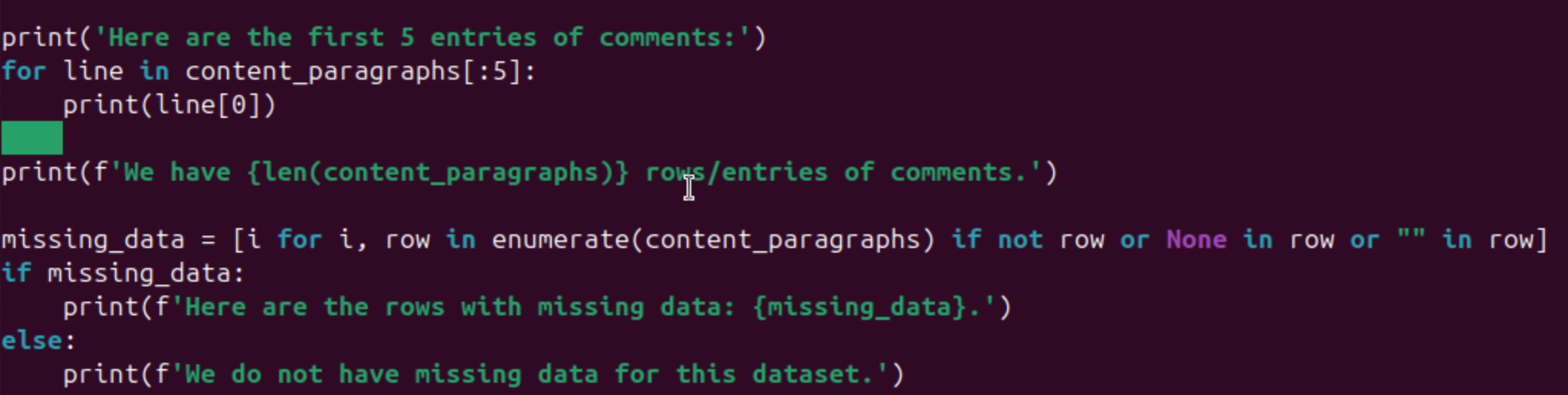
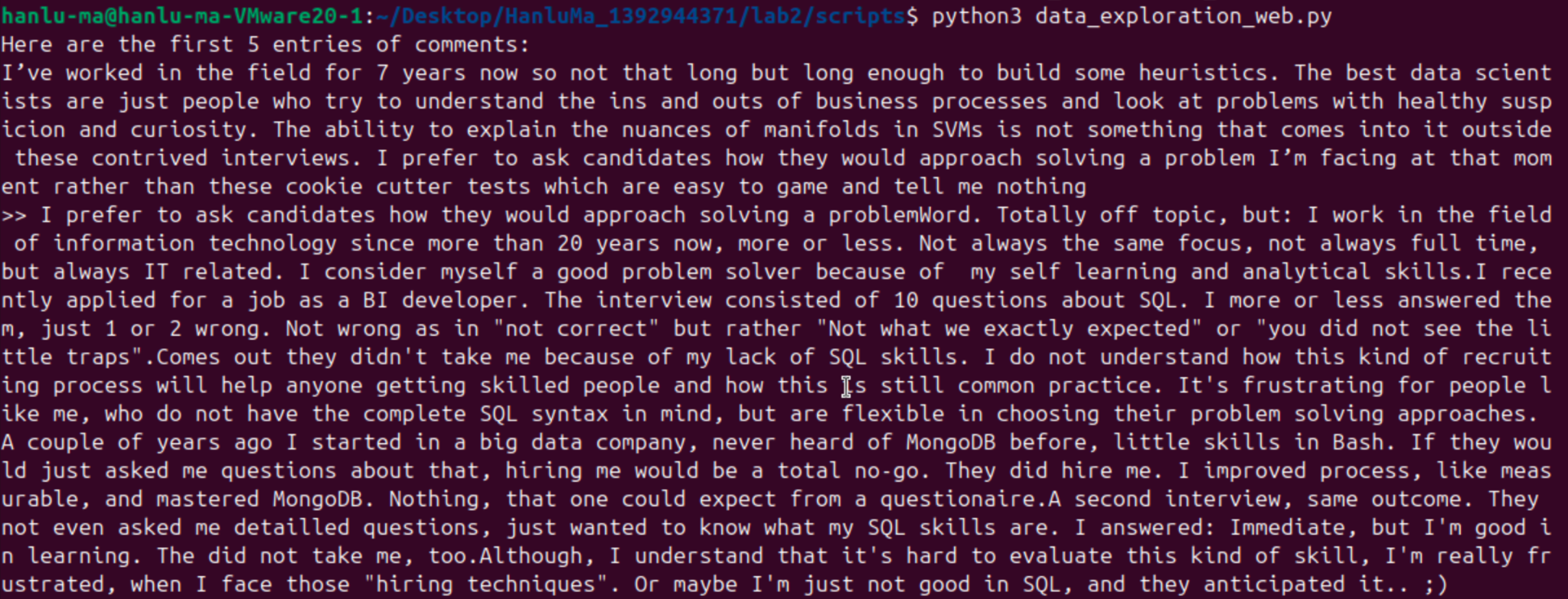
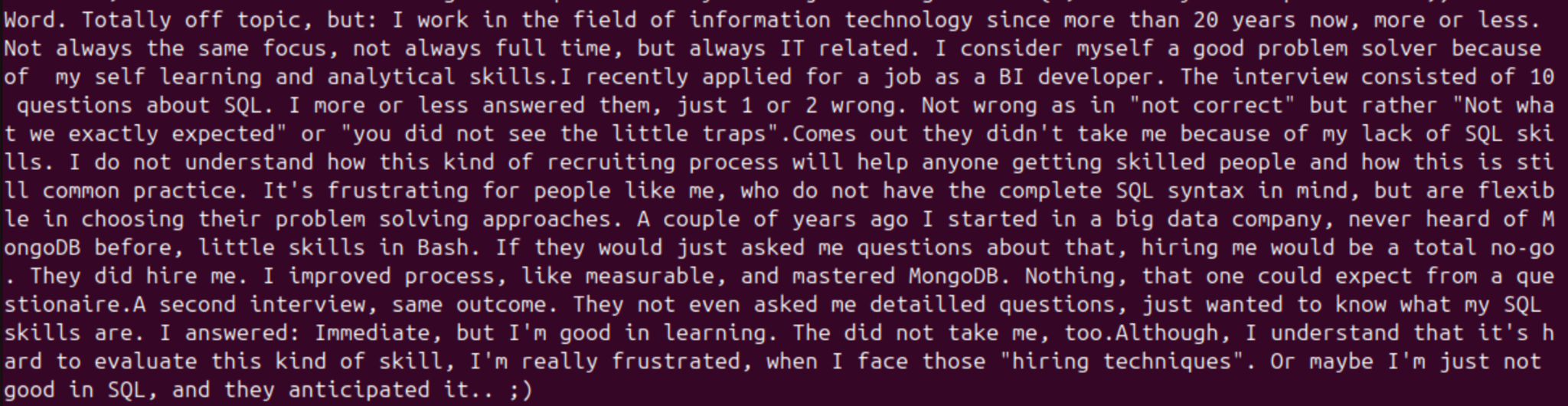
Data Science Professional Practicum (DSCI 560)

Laboratory Assignment 2

1. Team Formation
   1. Team name:
   2. Team numbers and their USC IDs:
      1. Hanlu Ma (USC ID: 1392-9443-71)
      2. Zhenyu Chen (USC ID: 2242-3773-15)
2. Data selection, Search, Find, and Collect
   1. Domain: job seeking content
   2. Reason
   3. Dataset links and descriptions
      1. ASCII Texts like Forum Postings and HTML
         1. Link: <https://news.ycombinator.com/item?id=24460141>
         2. Description: this webpage entails peoples’ responses about how to prepare for data scientist interviews. After the data exploration phase, this dataset will provide some ‘empirical’ and ‘humane’ advice and suggestions about data scientist interview preparation.
3. Data collection
   1. CSV or Excel
   2. ASCII Texts like Forum Postings and HTML
      1. Code  
           
         
      2. Output  
         For data exploration, we generate the first 5 entries of comments, the number of comments/rows/entries, and the missing data situation to check the validity of the dataset.  
           
           
         
      3. Rationale of Code
         1. We use the ‘get’ method from the requests library to get the link and utilize ‘BeautifulSoup’ to obtain the information. Then we locate the html location of the context we want to parse, which is ‘commtext c00’ in this case. After deep diving into the webpage html, we notice that some comments go along with the div, but others appear with ‘p’ or ‘i’ tag under the ‘commtext c00’ div. Therefore, we use the ‘final\_all’ method and the ‘get\_text’ method to both obtain information along with the div and within the category of the corresponding div. After obtaining the relevant information and storing in the relevant variable ‘content\_paragraphs’, we write it into csv under the column name of ‘Comments’. Lastly, we have few lines to do some sanity checks about the dataset in terms of dataset length, missing values, and sample data extraction.
   3. PDF and Word Documents that require conversion and OCR